[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
G06F 17/30

[12]发明专利申请公开说明书

[21]申请号 99804913.1

[43]公开日 2001年5月23日

[11]公开号 CN 1296589A

[22]申请日 1999.3.12 [21]申请号 99804913.1
[30]优先权
[32]1998.4.10 [33]US [31]09/058,635
[86]国际申请 PCT/GB99/00752 1999.3.12
[87]国际公布 WO99/53418 英 1999.10.21
[85]进入国家阶段日期 2000.10.9
[71]申请人 国际商业机器公司
地址 美国纽约
[72]发明人 索门·查卡雷贝蒂
拜伦·E·多姆

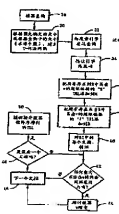
[74]专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 于静

权利要求书7页 说明书10页 附图页数5页

[54]发明名称 通过超级链接扩散特性

[57]摘要

响应查询根据普及性对广域网计算机网络(如万维网)页面进行分级的系统和方法。再有,使用查询和来自搜索引擎的对查询的响应,该系统和方法找出附加的为好的扩展的搜索术语的关键词,尤其在查询时刻实时产生一个本地同义词典。



ISSN1008-4274

00·10·09

权 利 要 求 书

1.一个包括数据存储装置的计算机（12），该数据存储装置包括一计算机可用介质（19，22），该介质中有计算机可用代码装置，用于响应查询以对一文档集合中的文档进行分级，该计算机可用代码装置有：

用于在第一文档中识别出对第二文档的参考的计算机可读代码装置；

用于接收一词汇间隔的计算机可读代码装置（30），该词汇间隔定义文档术语个数；

用于接收包括一个或多个查询术语的查询的计算机可读代码装置（28）；以及

用于确定在第一文档中在指向第二文档的参考的词汇间隔范围内存在至少一个查询术语的次数，并据此对文档分级的计算机可读代码装置（40、42、44、46、48）。

2.如权利要求 1 中申明的计算机（12），这里的文档可通过广域计算机网络访问，而且参考包括一个统一资源列表（URL）。

3.如权利要求 2 中申明的计算机（12），这里的词汇间隔是根据查询建立的。

4.如权利要求 2 中申明的计算机，进一步包含计算机可读代码装置（50）用于根据文档中查询术语各自出现在参考的词汇间隔范围内的次数来对多个文档分级。

5.如权利要求 2 中申明的计算机（12），进一步包含：

用于接收文档集合“U”的计算机可读代码装置（52）；

对于集合“U”中的至少一个测试文档“u”，把集合“U”中包括至少一个对测试文档“u”的参考的那些文档定义为近邻文档“N(u)”的计算机可读代码装置（70）；

对于至少一个近邻文档“N(u)”中的至少一个文档术语，确定是否该至少一个文档术语处在近邻文档“N(u)”中对测试文档“u”的一个参考的预先确定间隔内的计算机可读代码装置（76，78）；以及

响应确定是否该至少一个文档术语处在一参考的预先确定间隔范围内的装置，从而输出一个信号的计算机可读代码装置。

6.如权利要求 5 中申明的计算机，这里当至少一个文档术语处在对测试文档“u”的参考的预先确定间隔范围内时，该输出装置使与这至少一个文档术语关联的计数器增量。

7.如权利要求 2 中申明的计算机，进一步包含：

响应包括一个或多个查询术语的查询，接收文档集合“U”的计算机可读代码装置，其中每个文档包含一个或多个文档术语；以及

当至少一个文档术语和对至少一个第一文档的参考二者在这至少一个文档的一个查询术语的预先确定间隔范围内时，在这至少一个第一文档和这至少一个文档术语之间定义相关性的计算机可读代码装置。

8.如权利要求 7 中申明的计算机，这里的相关性与一权重关联，该权重是基于该文档术语和对第一文档的参考处在文档集合“U”中一查询术语的预先确定间隔范围内的次数。

9.一个计算机程序装置，包含：

可由数字处理装置读出的计算机程序存储器装置；以及

在程序存储器装置上的程序装置，它包括可由数字处理装置执行的指令，用于完成在一文档集合中找出关键词的方法步骤，这些方法步骤包含：

接收该文档集合；

确定该文档集合中的参考文档和被参考文档，参考文档是该集合中的含有对被参考文档的参考的那些文档；

对参考文档中的每个文档术语，确定该文档术语出现在对被参考文档的参考的预先确定间隔范围内的次数；以及

根据各次的次数对这些文档中的至少一些文档术语进行分级。

10.如权利要求 9 中申明的计算机程序装置，这里的方法步骤进一步包含：每当一文档术语出现在对一被参考文档的参考的预先确定隔范围内时，使相应的计数器增量。

11.如权利要求 9 中申明的计算机程序装置，这里的方法步骤进一步包含:

响应包括一个或多个查询术语的查询，接收文档集合“U”；以及

定义至少一个第一文档和至少一个文档术语之间的相关性，如果该文档术语和对该第一文档的一个参考二者处在一查询术语的预先确定间隔范围内的话。

12.如权利要求 11 中申明的计算机程序装置，这里的相关性与一权重关联，该权重基于该文档术语和对第一文档的参考处在文档集合“U”中一查询术语的预先确定间隔范围内的次数。

13.如权利要求 9 中申明的计算机程序装置，这里的方法步骤进一步包含:

接收一词汇间隔，该词汇间隔定义一文档术语个数;

接收包括一个或多个查询术语的查询；以及

确定至少一个查询术语在第一文档中出现在对第二文档的参考的词汇间隔范围内的次数，并据此对文档分级。

14.如权利要求 13 中申明的计算机程序装置，这里的文档可通过广域计算机网络访问，而且该参考包括一个统一资源列表（URL）.

15.如权利要求 14 中申明的计算机程序装置，这里的词汇间隔是根据查询建立的。

16.如权利要求 14 中申明的计算机程序装置，这里的方法步骤进一步包含：根据文档中查询术语出现在参考的词汇间隔范围内的次数来对多个文档分级。

17.在计算机存储的文档中找出文档术语和由一个或多个查询术语代表的查询主题之间的关联的方法，这些文档有各自的文档名，该方法包含如下步骤:

响应查询术语，接收至少一个文档列表；以及

当在一文档中在一查询术语的预先确定间隔范围内找出一文档术语和一文档名二者时，输出一个信号，以代表该文档术语和该查询主题之间的关联。

18.如权利要求 17 中申明的方法，这里的输出步骤包括：

构成一个有顶点的双枝图，其顶点代表在接收步骤接收的文档；以及

对至少一些文档(u)，和对每个文档(u)中的至少一些文档术语(t)，当在一查询术语的预先确定间隔范围内找到该文档术语(t)和文档(u)的文档名二者时，对双枝图中的边缘(b,u)的权重增量，这里边缘(t,u)代表文档(u)和文档术语(t).

19.如权利要求 17 中申明的方法，进一步包含如下步骤：

在文档列表中确定参考文档和被参考文档，参考文档是列表中含有对被参考文档的参考的文档；

对参考文档中的每个文档术语，确定该文档术语出现在对一被参考文档的参考的预先确定的间隔范围内的次数；以及

根据各自的次数，对这些文档中的至少一些文档术语进行分级。

20.如权利要求 17 中申明的方法，进一步包含如下步骤：

接收一词汇间隔，该词汇间隔定义一文档术语个数；

确定至少一个查询术语在第一文档中出现在对第二文档的参考的词汇间隔范围内的次数，并据此对文档分级。

21.如权利要求 20 中申明的方法，这里的文档可通过广域计算机网络访问，而且该参考包括一个统一资源列表（URL）.

22.如权利要求 20 中申明的方法，这里的词汇间隔是根据查询术语建立的.

23.如权利要求 20 中申明的方法，进一步包含根据文档中查询术语出现在参考的词汇间隔范围内的各自次数来对多个文档分级的步骤.

24.一个包括数据存储装置的计算机，该数据存储装置包括一计算机可用介质，该介质中有计算机可用代码装置，用于在计算机存储的文档中找出文档术语和由一个或多个查询术语代表的查询主题之间的关联，这些文档有各自的文档名，该计算机可用代码装置有：

用于响应查询术语，接收至少一个文档列表的计算机可读代码装

置；以及

当在一文档中在一查询术语的预先确定间隔范围内找出一文档术语和一文档名二者时，用于输出一个代表该文档术语和该查询主题之间关联的信号的计算机可读代码装置。

25.如权利要求 24 中申明的计算机，这里的输出装置包括：

构成一个有顶点的双枝图的计算机可读代码装置，各顶点代表在接收步骤接收的文档；

对至少一些文档(u)，和对每个文档(u)中的至少一些文档术语(t)，当在一查询术语的预先确定间隔范围内找到该文档术语(t)和文档(u)的文档名二者时，用于对双枝图中的边缘(t,u)的权重进行增量的计算机可读代码装置，这里边缘(t,u)代表文档(u)和文档术语(t)。

26.如权利要求 25 中申明的计算机，进一步包含：

在文档列表中确定参考文档和被参考文档的计算机可读代码装置，参考文档是列表中含有对被参考文档的参考的文档；

对参考文档中的每个文档术语，确定该文档术语出现在对一被参考文档的参考的预先确定间隔范围内的次数的计算机可读代码装置；

根据各自的次数，对这些文档中的至少一些文档术语进行分级的计算机可读代码装置。

27.如权利要求 26 中申明的计算机，进一步包含：

用于接收一词汇间隔的计算机可读代码装置，该词汇间隔定义一文档术语个数；以及

确定至少一个查询术语在第一文档中出现在对第二文档的参考的词汇间隔范围内的次数，并据此对文档分级的计算机可读代码装置。

28.如权利要求 27 中申明的计算机，这里的文档可通过广域计算机网络访问，而且该参考包括一个统一资源列表（URL）。

29.一个包括数据存储装置的计算机，该数据存储装置包括一个计算机可用介质，该介质中有计算机可用代码装置，用于在一文档集合中找出关键词，该计算机可用代码装置有：

用于接收文档集合的计算机可读代码装置；

确定该文档集合中的参考文档和被参考文档的计算机可读代码装置，其参考文档是该集合中的那些包含对被参考文档的参考的文档；

对参考文档中的每个文档术语，确定该文档术语出现在对一被参考文档的参考的预先确定的间隔范围内的次数的计算机可读代码装置；以及

根据各自的次数，对这些文档中的至少一些文档术语进行分级的计算机可读代码装置。

30.如权利要求 29 中申明的计算机，进一步包含计算机可读代码装置用于每当一文档术语出现在对一被参考文档的参考的预先确定间隔范围内时使相应的计数器增量。

31.如权利要求 30 中申明的计算机，进一步包含：

响应包括一个或多个查询术语的查询，接收文档集合“U”的计算机可读代码装置；以及

当至少一个文档术语和对至少一个第一文档的参考二者在一个查询术语的预先确定间隔范围内时，在这至少一个第一文档和这至少一个文档术语之间定义相关性的计算机可读代码装置。

32.如权利要求 31 中申明的计算机，这里的相关性与一权重关联，该权重是基于该文档术语和对第一文档的参考处在文档集合“U”中一查询术语的预先确定间隔范围内的次数。

33.如权利要求 32 中申明的计算机，进一步包含：

接收一词汇间隔的计算机可读代码装置，该词汇间隔定义一文档术语个数；

接收包括一个或多个查询术语的查询的计算机可读代码装置；以及

确定至少一个查询术语在第一文档中出现在对第二文档的参考的词汇间隔范围内的次数，并据此对文档分级的计算机可读代码装置。

34.如权利要求 33 中申明的计算机，这里的文档可通过广域计算机网络访问，而且该参考包括一个统一资源列表（URL）。

35.一个包括数据存储装置的计算机，该数据存储装置包括一计算

机可用介质，该介质中有计算机可用代码装置用于响应一查询对一文档集合中的文档进行分级，该计算机可用代码装置有用于接收文档集合“U”的计算机可读代码装置；

对于集合“U”中的至少一个测试文档“u”，把集合“U”中包括至少一个对测试文档“u”的参考的那些文档定义为近邻文档“N(u)”的计算机可读代码装置；

对于至少一个近邻文档“N(u)”中的至少一个文档术语，确定是否该至少一个文档术语处在测试文档“u”的近邻文档“N(u)”的一个参考的预先确定间隔内的计算机可读代码装置；以及

响应确定是否该至少一个文档术语处在一参考的预先确定间隔范围内的装置，从而输出一个信号的计算机可读代码装置。

36.如权利要求 35 中申明的计算机，这里当该至少一个文档术语处在对测试文档“u”的参考的预先确定间隔范围内时，该输出装置使与这至少一个文档术语关联的计数器增量。

说　　明　　书

通过超级链接扩散特性

一般地说，本发明涉及信息检索，更具体地说，涉及在例如万维网上高效率地和有效果地检索超文本文档(document)的方法和装置。

称作因特网的广域计算机网络，特别是称作万维网的因特网部分，使用户能访问大量信息。毫不惊奇，已经提供了若干个搜索引擎，用户能向其中输入查询，而搜索引擎能使用各种方案返回万维网站清单以响应这些查询，从而便于从万维网挖掘信息。这些万维网站一般代表由计算机存储的文档，用户能访问这些文档以得到关于该特定站点主题的信息。

通常，与大多数计算机搜索方法相似，万维网搜索引擎使用某种关键词搜索策略，其中，用户输入查询的一个或多个术语以某种方式与万维网文档中的术语进行匹配，以向查询用户返回一个特定万维网站清单。然而，发生的情况是大多数查询的长度只有一至三个词，这样，通常这些查询的范围很广。这意味着有大量万维网站可能含有一个查询的一个或多个词，而且，如果搜索引擎返回所有可能的候选者，那么用户可能需要筛选成百或数千个文档。

再有，可能发生这样的情况，即响应一个查询时，那些最贴近该查询的万维网站可能根本未被返回。更具体地说，一个查询使用的术语可能在最贴近该查询的万维网站中不出现。例如，在为当今最普及的两个浏览器的万维网站中根本没出现“浏览器”这个术语。相反，这些万维网站使用“浏览器”以外的其他词来说明这些网站的主题。结果，如果一个用户向使用简单的关键词查询策略的搜索引擎输入词“浏览器”，那么这些网站将不会被返回给用户。

然而，如本发明认识到的那样，因特网用户不知不觉地在合作搜索、阅读、评论和判断万维网文档的质量。这种合作大部分通过万维网页的汇编反映在大部分（如果不是全部）万维网页中，这些万维网

页通常描述和指向那些被看作是高质量的其他网页。

更具体地说，一个万维网页以超级链接的形式指向其他万维网页，这实质上是在第一文档（即第一万维网页）中参考其他文档（即其他网页）。超级链接使用户能通过利用计算机鼠标或其他指向与点击装置“点击”该超级链接，从而选择立即访问另一个万维网页。如这里所认识到的那样，这种参考万维网页可以是这样一些术语的丰富来源，这些术语已经广泛地与那些被参考万维网页关联，即使那些被参考网页本身并不使用这些术语。结果，这些术语能被用于改善万维网搜索查询结果。本发明进一步认识到，通过对一文档的参考（例如一个超级链接）来有效地扩散特性的这些原理不仅适用于万维网，也能应用于被链接如专利、学术论文、文章、书籍、电子邮件等文档的任何实体。

因此，本发明的一个目的是提供一种通过超级链接扩散特性的方法和系统。本发明的另一目的是提供一种方法和系统，用于响应用户查询，在一组文档中对文档进行分级。本发明的又一目的是提供一种方法和系统，用于在一组文档中找出关键词。本发明的再一个目的是提供一种方法和系统，用于在计算机存储的文档中找出文档术语和由一个或多个查询术语所代表的查询主题之间的关联。本发明的另一目的是提供一种用于万维网搜索的方法和系统，这种万维网搜索便于使用而且节省费用。

本发明是根据这里所发明的步骤进行编程的通用计算机，以响应查询对一组文档中的文档进行分级。本发明还能实现为一个制造的物品－机器部件－它被一数字处理系统所用并且有形地实现一个指令程序，该程序可由该数据处理装置执行，以在计算机存储的文档中找出文档术语和查询主题之间的关联。本发明在一个关键机器部件中实现，该部件使一数字处理装置完成这里发明的方法步骤。

根据本发明，该计算机包括计算机可读代码装置，用于在第一文档中识别出对第二文档的参考。计算机可读代码装置接收一个定义文档术语个数的词汇间隔(lexical distance)。再有，该计算机包括计算机

可读代码装置用于接收包括一个或多个查询术语的查询，以及计算机可读代码装置用于确定在第一文档中出现的位于对第二文档进行参考的词汇间隔范围内的至少一个查询术语的次数，用于据此对文档分级。

在一个实施例中，可通过广域计算机网络访问文档，而参考包括一个统一资源列表（URL）。如果希望的话，根据查询建立词汇间隔。

最好是该计算机还包括计算机可读代码装置用于根据在文档中参考的词汇间隔范围内存在查询术语的相应次数来对多个文档进行分级。此外，该计算机包括计算机可读代码用于接收文档集合“U”。提供的计算机可读代码用于对集合“U”中的至少一个测试文档“u”，把集合“U”中包括至少一个对测试文档“u”的参考的那些文档定义为邻居文档“N(u)”。再有，对于至少一个邻居文档“N(u)”中的至少一个文档术语，计算机可读代码装置确定是否至少有一个文档术语处在测试文档“u”的邻居文档“N(u)”中的一个参考的预先确定的间隔内（即在一预先确定的术语个数范围内）。根据本发明，计算机可读代码装置于是发出一个信号，以响应确定是否至少有一个文档术语处在一个参考的预先确定间隔范围内的那个装置。当这至少一个文档术语处在对测试文档“u”的一个参考的预先确定间隔范围内时，该输出装置使一个与这至少一个文档术语相关联的计数器增量。

除了上面概述的逻辑外，该计算机还能包括计算机可读代码装置用于响应包括一个或多个查询术语的查询，从而接收一个文档集合“U”，其每个文档包含一个或多个文档术语。提供的计算机可读代码装置用于确定在至少一个第一文档和至少一个第一文档术语之间的相关性，如果该第一文档术语和对该第一文档的一个参考二者都处在一个查询术语的一个预先确定的间隔范围内的话。如果希望，该相关性与一权重相关联，而该权重是基于文档集合“U”中第一文档术语和对第一文档的一个参考二者处在一查询术语的一个预先确定间隔范围内的次数。

在另一方面，一个计算机程序装置包括一个可由数字处理装置读出的计算机程序存储装置，以及程序存储装置上的一个程序装置，它

包括可由该数字处理装置执行的指令，用于完成在一文档集合中找出关键词的方法步骤。这些方法步骤包括接收一组文档，然后确定该组文档中的参考文档和被参考文档，参考文档是该组中含有对被参考文档的参考的那些文档。对参考文档中的每个文档术语，确定该文档术语出现在对一被参考文档的参考的预先确定间隔范围内的次数。根据各相应的次数，在这些文档中至少有一些文档术语被分级。还公开了一种计算机，用于实现上面描述的程序装置。

在又一方面，公开了一种由计算机实现的方法，用于在由计算机存储的文档中找出在文档术语和由一个或多个查询术语代表的查询主题之间的关联。根据本发明，这些文档各有相应的文档名。该方法包括响应查询术语接收至少一个文档列表，然后，当在一文档中在一查询术语的一个预先确定间隔范围内找出一文档术语和一文档名二者时，输出一个信号代表该文档术语和该查询主题之间的关联。还公开了一个执行上面概括的方法的计算机。

在另一方面，一个计算机包括一个数据存储装置，该数据存储装置又包括一个计算机可用介质，它有计算机可用代码装置用于响应一查询对一个文档集合中的文档进行分级。该计算机可用代码装置有计算机可读代码装置用于接收一个文档集合“U”，以及计算机可读代码装置用于对集合“U”中的至少一个测试文档“u”，把集合“U”中包括至少一个对测试文档“u”的参考的那些文档定义为邻居文档“N(u)”。此外，对至少一个邻居文档“N(u)”中的至少一个文档术语，计算机可读代码装置确定该文档术语是否处在该测试文档“u”的邻居文档“N(u)”内的一个参考的预先确定间隔范围内。然后，计算机可读代码装置输出一个信号，以响应该确定装置。

现在将参考附图，仅以举例方式描述本发明，这些附图中：

图 1 是通过超级链接扩散文档特性的本计算机系统略图；

图 2 是计算机程序产品略图；

图 3 是一逻辑的流程图，该逻辑用于响应一查询，增长已提供的万维网站列表；

图 4 是一逻辑的流程图，该逻辑用于响应一查询，从所产生的页面列表中回送“高质量”页面；

图 5 是一流程图，所显示的逻辑用于通过超级链接找出描述性术语（这里也称作特性）；以及

图 6 是一流程图，所显示的逻辑用于在计算机存储的文档中找出文档术语和由一个或多个查询术语代表的查询主题之间的关联。

发明详述

首先参考图 1，图中显示通过超级链接找出描述性术语的系统，总体用 10 表示。在所示具体结构中，系统 10 包括一数字处理装置，如计算机 12。在一个预定的实施例中，计算机 12 可以是图中所示由 ArmonK, N.Y.的国际商用机器公司(IBM)制造的个人计算机，或者计算机 12 可以是任何计算机，包括以诸如 AS/100 为商标出售的计算机，并伴有 IBM 网络工作站。或者，计算机 12 可以是一个 Unix 计算机，或 OS/2 服务器，或 Windows NT 服务器，或运行 AIX3.2.5 的带有 128 MB 主存储器的 IBM RS/6000 250 工作站，或 IBM 膝上计算机。（UNIX 是 Open Group 的商标，AS/400、OS/2、RS/6000 和 AIX 是国际商用机器公司的商标，Windos NT 是微软公司的商标）。

计算机 12 访问一个因特网搜索引擎 14。在一个实施例中，该搜索引擎 14 是由 Alta Vista 制造的，可以理解，其他搜索引擎当然也可使用。搜索引擎 14 从计算机 12 接收查询，并响应该查询向计算机 12 回送一个计算机存储文档列表，更具体地回送一个万维网站列表 16，利用这一列表，计算机 12 能经由称作万维网 18 的因特网部分进行通信。

此外，计算机 12 包括一个特性扩散器模块 19，它作为一系列计算机可执行的指令，由计算机 12 内的处理器来执行。这些指令可以驻留在例如计算机 12 的 RAM 中。这里的这些流程图说明本发明的模块 19 承担在计算机程序软件中实现的编程指令结构。本领域技术人员将会理解，这些流程图说明根据本发明实现其功能的逻辑单元（如计算机程序代码单元或电子逻辑电路）的结构。显然地，本发明在其基本

实施例中是由机器部件实现的，这些机器部件执行逻辑单元，其执行方式是指示数字处理装置（即计算机）完成与所示逻辑单元对应的一系列功能步骤。

换言之，模块 19 可以是一计算机程序，它作为一系列计算机可执行的指令由计算机 12 内的处理器执行。

另一种作法是，这些指令可以存储在具有计算机可读介质的数据存储装置上，如图 2 中所示软盘 20。软盘 20 可包括计算机可用介质 22，它电子存储计算机可读程序代码单元 A－D。或者，这些指令可存储在 DASD 阵列，磁带、传统的硬盘驱动器、电子只读存储器、光存储装置、或其他适当的数据存储装置上。在本发明的一个示例实施例中，计算机可执行指令可以是编译的 C⁺⁺兼容代码行或超文本标记语言（HTML）兼容代码行。

图 1 还显示出系统 10 可包括本领域公知的外围计算机设备，包括输入装置，如计算机键盘 24 和/或计算机鼠标 25。可使用除图中所示以外的其他输入装置，如跟踪球、小键盘（Keypad）、触模屏、以及语音识别装置。还提供了一个输出装置，如视频监视器 26。其他输出装置可以使用，如打印机和其他计算机等。

现在参考图 3，图中可看到由模块 19 承担的第一过程（这里称作“过程 A”）的逻辑。在块 28 开始，接收一个可能由键盘 24 输入的用户查询。用户查询由一个或多个查询术语组成，如“高山(high mountains)”。

进到块 30，建立了词汇间隔“1”。在一个最佳实施例中，为便于简短地说明，该词汇间隔“1”用术语的整数个数定义一个窗口。词汇间隔“1”可有固定值，或者，作为另一种方式，词汇间隔“1”的值可根据查询中的中个数来建立。例如，词汇间隔“1”的值可与查询术语个数成反比。

在块 32，该查询被送到搜索引擎 14。根据搜索引擎原理，搜索引擎 14 返回一个符合该查询的万维网站列表 16，该列表在一结果集合“R”中返回，并在块 34 收到该结果集合“R”。通常，结果集合是作

为万维网站名列表返回的，这些万维网站名被称作统一资源定位符（URL）。

进到块 36，于是该逻辑按下述方式扩展结果集合“R”。首先，把与结果集合“R”中的一个或多个元素“r”存在超级链接的所有“s”URL 加到结果集合“R”中。这样，在块 36，在第一文档中识别出了对第二文档的参考。

接下来，在块 38，把全部“t”URL 加到结果集合“R”中，这“t”URL 的特征是存在从结果集合“R”中的任何元素“r”到该 URL 的超级链接，这一扩展了的集合表示为“S”。这样，在块 36 和 38，通过把由结果集合“R”中的 URL 中的超级链接所指向的那些 URL 或者借助超级链接指向 R 中 URL 的那些 URL 添加到结果集合“R”中，使结果集合“R”扩展为扩展集合“S”。

该逻辑从块 38 移到块 40，从而进入对扩展集合“S”中每个文档的“DO”循环。在决策菱形 42 中，确定是否在该文档的 URL“u”的词汇间隔范围内出现任何查询术语，即是否在被测试文档中有任何查询术语出现在指向扩展集合“S”中第 u 个文档的超级链接的词汇间隔范围内。如果是，则在块 44 使与扩展集合“S”中的第 u 个文档关联的一个计数器 u 增加 1，然后在块 46 检索出下一个文档。这样，该逻辑确定了在第一文档中有至少一个查询术语存在于对第二文档的参考的词汇间隔范围内的次数，用于据此对文档分级（见下述）。

如果在决策菱形 42 处的测试是否定的，则逻辑直接移到块 46。逻辑从块 46 移到决策菱形 48，以确定是否已完成“DO”循环，如果否，则逻辑循环回到决策菱形 42。另一方面，一旦完成了“DO”循环，过程则移到块 50，返回一个按计数器值递减顺序的 URL 有序集合。

现在参考图 4，可理解“B”过程，它试图根据这里某些术语的重要性对从过程“A”返回的头“N”个 URL 进行重新排序。在块 52 开始，接收一组文档。这组文档可以是例如在块 50 输出的头“N”个（例如 20 个）URL。对于这一组文档，进入一个“DO”循环，并在块 54

把索引变量“v”设为等于被测试的 URL.

移到块 56，在那里确定引用被测试 URL“v”（例如，通过包含一个指向被测试 URL“v”的超级链接）的所有 URL“u”（或其子集）. 接下来，进到块 58，在那里检索出指向被测试 URL“v”的超级链接所属的那些 URL 中的所有锚入文本.

“锚入文本”是指一个文本中直接与一超级链接或者其他参考或引用相关联的文本。例如，在这样一段话“One of the earliest high-energy nuclear accelerators was built at(最高的高能核子加速器之一曾建在)<A HREF= http://www. CERN. ch>CERN, the European Laboratory for Particle Physics(欧洲粒子物理实验室)</A>，中，超链接是短语 http://www. CERN. ch，而锚入文本是介于“<A>…</A>”之间的材料。利用此例，对于例如为 5 的词汇间隔，在该锚入文本的词汇间隔范围内的术语是“nuclear accelerators was built at”，而不在该锚入文本的词汇间隔范围内的术语是“One of the earliest high-energy”.

然后，对每个查询术语，在块 60 进入一个嵌入的“DO”循环。进入决策菱形 62，确定被测试查询术语在被测试文档中出现的频度是否大于锚入文本的某参考集合中的参考频度，如由各种传统的统计技术之一所确定的那样.

当被测试文档中的被测试查询术语出现的频度大于参考频度时，该过程移到块 64，在那里把该被测试文档标注为重要的. 否则，被测试文档不被标注为重要的. 在每种情况中，每个文档都可与一个计数器或其他值相关联，该计数器或其他值代表由上述测试得到的它的重要性. 在上面讨论的“DO”循环结束时，这头“N”个 URL 按其重要性排序.

现在参考图 5，图中显示通过超级链接找出描述性术语的过程“C”. 在块 68 处开始，接收 URL“u”的一个集合“U”，并对集合“U”中的每个单个 URL“u”，进入一个“DO”循环. 在块 70，确定 URL“u”的近邻“N(u)”集合.“近邻(in-neighbour)”是指 URL

集合“U”中的含有指向被测试文档“u”的超级链接的文档。换一种说法，可把近邻集合N(u)看作是参考被参考文档“u”的参考文档。

对近邻集合N(u)的每个元素（即文档术语），在块72进入一个嵌入的“DO”循环。移到块74，一个计数器与近邻集合N(u)的每个术语关联。接着，进入一个双嵌入“DO”循环进入决策菱形76，确定被测试术语是否在指向被测试文档“u”的一个参考（例如超级链接）的预先确定间隔范围内。这个预先确定间隔可以是上文讨论的词汇间隔。如果被测试术语处在指向被测试文档的一个参考的预先确定间隔范围内，则在块78对计数器加1。否则，该计数器不加1。当文档集合“U”中所有文档“u”的近邻集合N(u)中的所有近邻的所有术语都已按上述作法测试过时，该逻辑移到块80，按各术语各自的计数器值对这些术语排序，并返回一个排序列表。

如本发明认识到的那样，在块80处的输出是文档集合“U”中术语的分级列表。这一分级列表能用于向用户建议额外的查询术语。而且，它可作为运行中的关联词典。此外，在块80处的输出能用于注释被超级链接接的文档集丛和术语集丛，作为许多搜索引擎的一个后处理步骤。

图6显示过程“D”的逻辑，用于找出计算机存储文档中文档术语和由一个或多个查询术语代表的查询主题之间的关联。在块82处开始，接收一个查询“Q”。该查询“Q”由一个或多个查询术语“q”构成。

在块84处，该查询被送到一搜索引擎，作为响应，从搜索引擎回收到一个文档列表。移到块86，在此处构成一个双枝图G=((T,U),E)，其顶点是在块84处返回的术语（T）和文档（U），这里T和U分别代表双枝图的文档术语分支和URL分支，而这里的E代表分枝之间的边缘。

进到块88，对每个文档，进入一个“DO”循环。进到块90，该文档被扫描，寻找URL“u”和查询术语“q”。接下来进到块92，对于在查询术语“q”的一个预先确定间隔范围内找出的每个文档术语“t”和URL“u”，进入一个“DO”循环，其中在块94处对边缘(t,u)E的权重增1。利用这一逻辑，如果在一文档中在一查询术语的预先确定

间隔范围内找到一文档术语和一文档名或引用（以超级链接的形式）二者，则输出一个信号，它代表该文档术语和该查询主题之间的关联。

如果希望的话，该“DO”循环能包括进入块 96，在这里对边缘 E:ai,j 定义的矩阵 A 确定一个单值分解(SVD)，这里 ai,j 是从第 i 个术语到第 j 个 URL 的边缘的权重。如本领域众所周知，在块 96 处对 SVD 的确定有效地对 A 进行了因式分解：A=USV，这里 S 是含有 A 的奇异值的对角矩阵，而 U 和 V 是用于进行正交交换的正交矩阵。在本领域中称作隐伏语义检索(Latent Semantic Indexing, LSI)的技术，如在美国专利 4,839,853 号中公开的那种，可用于对全集进行预处理，特别是把文档－术语矩阵 A 分解为 USV，这里 U 给出从术语空间到可称作 LSI 或概念空间的线性投影。几百个 LSI 维数“k”足够了。

然而，LSI 搜索并不使用 U 矩阵，而本发明使用 U 矩阵，如下述。每个术语被映射到 LSI 空间，其每个文档由一个 K 维矢量序列代表。查询本身被变换成这种矢量的一个短序列。然后，这些文档被扫描，该逻辑试图使查询矢量与文档中的一个矢量小窗口匹配。如果存在一个低成本（即“好的”）匹配，则对附近的引用，即超级链接，投一个大的赞成票，可以用一种最小成本匹配策略来对成本进行估计，匹配与术语 t1 和 t2 对应的矢量所需的边缘成本就是它们在 U 中投影之间的距离。作为一例，查询“auto makers（汽车制造商）”可以以小成本匹配于文本序列“companies making cars(制造汽车的公司)”，于是对这种类似短语附近发生的引用投赞成票。

与 LSI 相反，本发明对每个文档保持一个 LSI 矢量序列。换言之，与 LSI 不同，本发明考虑匹配 LSI 矢量序列和使用评分对邻近的引用投票。

如果希望的话，该过程可在块 98 向用户返回建议的搜索术语。为确定这些建议的术语，该逻辑按降值顺序对在块 96 中确定的 SVD 左矢量（即“U”的第一列）上有投影的那些术语进行排序。然后，在块 98 将排序列表中的头“k”个术语返回，这里“k”是一个预先确定的整数，例如 5。

说 明 书 附 图

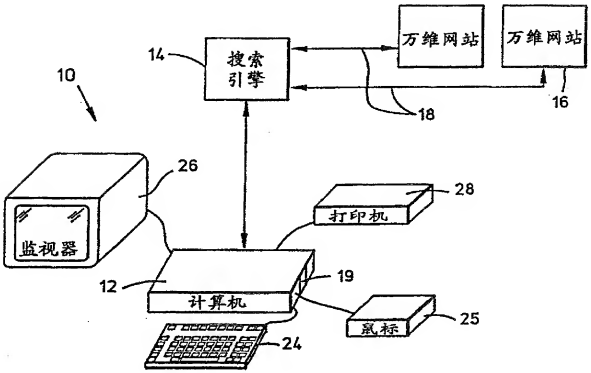


图1

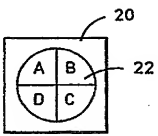


图2

图3

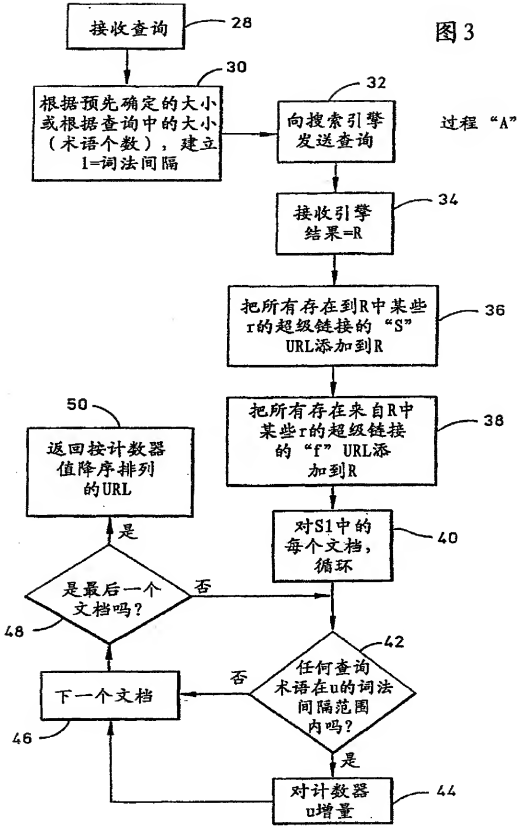
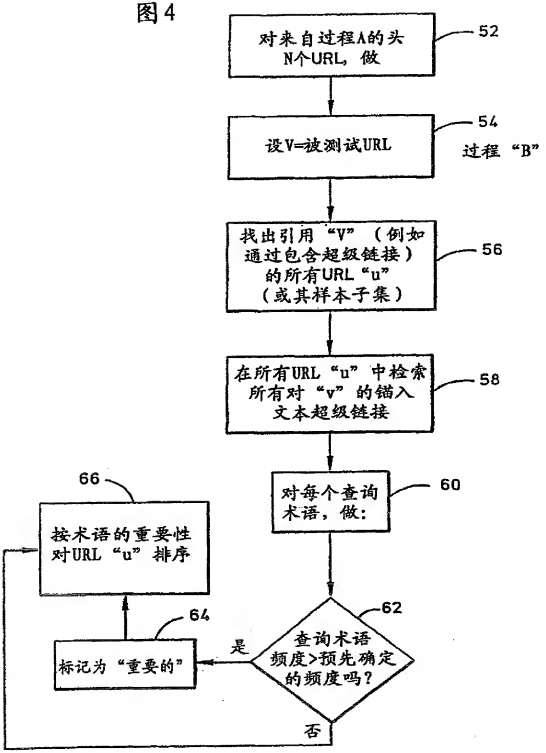

接收查询
28
30
根据预先确定的大小或根据查询中的大小（术语个数），建立1=词法间隔
32
向搜索引擎发送查询
过程 "A"
接收引擎结果=R
34
把所有存在到R中某些r的超级链接的 "S" URL添加到R
36
把所有存在来自R中某些r的超级链接的 "f" URL添加到R
38
对S1中的每个文档，循环
40
任何查询术语在u的词法间隔范围内吗？
42
否
是
对计数器u增量
44
下一个文档
46
否
是最后一个文档吗？
48
是
返回按计数器值降序排列的URL
50

图4


对来自过程A的头N个URL，做
52
设V=被测试URL
54
过程 "B"
找出引用 "V"（例如通过包含超级链接）的所有URL "u"（或其样本子集）
56
在所有URL "u" 中检索所有对 "v" 的锚入文本超级链接
58
对每个查询术语，做：
60
查询术语频度>预先确定的频度吗？
62
是
否
标记为 "重要的"
64
按术语的重要性对URL "u" 排序
66

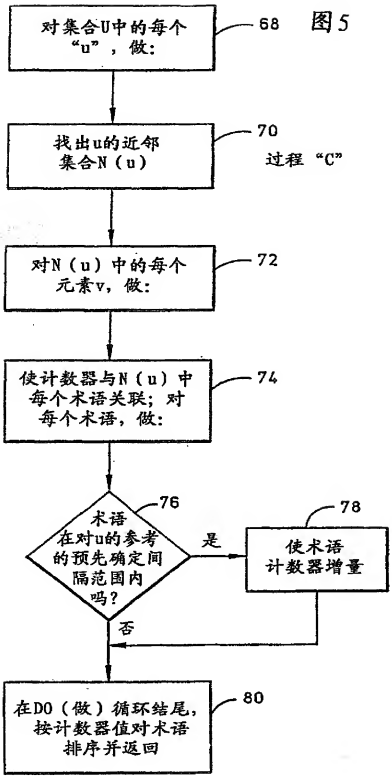
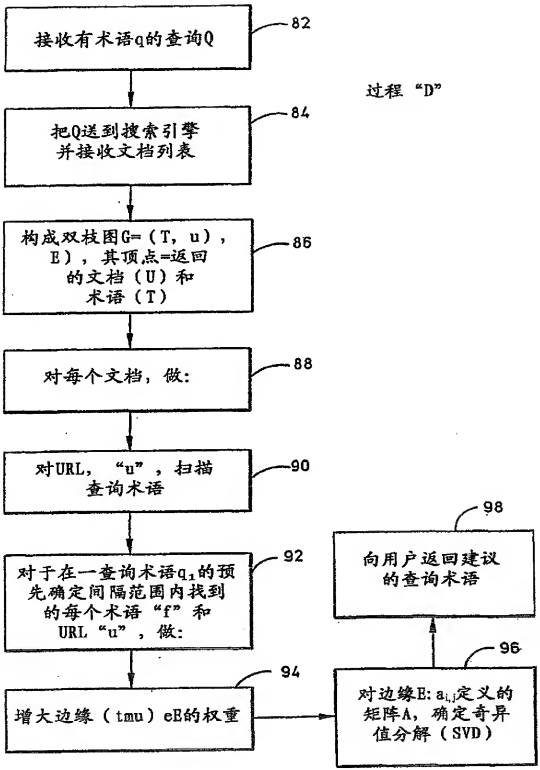

对集合U中的每个“u”，做：
68
找出u的近邻集合N（u）
70
过程“C”
对N（u）中的每个元素v，做：
72
使计数器与N（u）中每个术语关联；对每个术语，做：
74
术语在对u的参考的预先确定间隔范围内吗？
76
是
使术语计数器增量
78
否
在DO（做）循环结尾，按计数器值对术语排序并返回
80

图5

图6


过程 "D"
接收有术语q的查询Q
82
把Q送到搜索引擎
并接收文档列表
84
构成双枝图G=（T，u），
E），其顶点=返回
的文档（U）和
术语（T）
86
对每个文档，做：
88
对URL，"u"，扫描
查询术语
90
对于在一查询术语q_1的预
先确定间隔范围内找到
的每个术语"f"和
URL"u"，做：
92
增大边缘（tmu）eE的权重
94
对边缘E:$a_{i,j}$定义的
矩阵A，确定奇异
值分解（SVD）
96
向用户返回建议
的查询术语
98